Bayesian Decision Trees for Predicting Survival of Patients: A Study on the US National Trauma Data Bank

V. Schetinin, L. Jakaite, J. Jakaitis

Department of Computer Science and Technology, University of Bedfordshire, UK

W. Krzanowski

College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK

Abstract

Trauma and Injury Severity Score (TRISS) models have been developed for predicting the survival probability of injured patients the majority of which obtain up to three injuries in six body regions. Practitioners have noted that the accuracy of TRISS predictions is unacceptable for patients with a larger number of injuries. Moreover, the TRISS method is incapable of providing accurate estimates of predictive density of survival, that are required for calculating confidence intervals. In this paper we propose Bayesian inference for estimating the desired predictive density. The inference is based on decision tree models which split data along explanatory variables, that makes these models interpretable. The proposed method has outperformed the TRISS method in terms of accuracy of prediction on the cases recorded in the US National Trauma Data Bank. The developed method has been made available for evaluation purposes as a stand-alone application.

Keywords: Bayesian prediction, survival probability, Markov chain Monte Carlo, classification tree, trauma care.

1. Introduction

Probabilities of survival for patients alive on arrival at a hospital are calculated by using the Trauma and Injury Severity Score (TRISS) system [4, 3, 21, 12, 22]. TRISS predictions are based on a logistic regression model

that considers up to three most severe injuries a patient can obtain in six regions of the body such as head, face, chest, abdomen, and extremities. The TRISS model takes into account screening tests as explanatory variables which include: age, systolic blood pressure, respiratory rate, the severity scores of injuries as well as Glasgow coma scores and type of injury. The screening tests are performed on the patient's arrival at an emergency unit.

To assist the practitioners with predictions of patient's survival, a TRISS Calculator has been made available online [6]. The Calculator requires the user to input the Abbreviated Injury Scales for the six regions, as well as the systolic blood pressure, respiratory rate, Glasgow coma score, and age of a patient. The output is a predicted survival probability for blunt or penetrating type of an injury.

According to the TRISS method, the screening tests are used to form two aggregated predictors. The abbreviated injury scales form the Injury Severity Score (ISS) and the systolic blood pressure, respiratory rate, and Glasgow coma score form the Revised Trauma Score (RTS). Such aggregated predictors have revealed unexplained fluctuations over actual probabilities of survival [25, 3, 22, 1, 26].

The match between the actual survival and predicted probabilities is considered as *calibration*, and can be visualized as a *calibration curve*, see e.g. [17, 33]. The ideal calibration curve is a 45 degree line with zero intercept. It has been found that the calibration of the TRISS model significantly deviated from the ideal curve [22, 28].

The regression coefficients of the TRISS model have been calculated on the Major Trauma Outcome Study (MTOS) database in the 1980s [8] and updated in 1990s [9]. Since that time, the technologies in trauma care have been advanced, and the TRISS model was shown to be providing over-pessimistic predictions, and so there is the need to recalibrate the TRISS model, see e.g. [28]. Such a recalibration has been attempted on the US National Trauma Data Bank (NTDB), [12], and the updated coefficients were published in [34]. Similar attempts have been undertaken in [10, 23].

As a way of improving TRISS predictions, the recalibration can be efficient when a model is given appropriately. In practice such a model is difficult to identify from a given set of data, see e.g. [3]. A model can be fitted to data with the likelihood maximization over a model parameter space. However, except for trivial cases this method requires much effort to overcome the optimization problem caused by areas of low likelihood values, so that the desired improvement in TRISS predictions cannot be guaranteed

[14, 18, 2, 24].

Despite the problems, the accuracy of TRISS predictions has been found acceptable when the types and severities of patients injuries are typical. However, for cases with four or more injuries as well as with some atypical combinations of injuries, the accuracy could be improved [22].

Practitioners are interested not only in accurate predictions of survival probabilities but also in estimating the uncertainty in predictions. In general, estimates of uncertainty are required to minimize risks of mistaken decisions and, particularly, to calculate confidence intervals. These intervals can be accurately estimated when a predictive probability density is fully known, but the desired estimates cannot be provided within a concept such as TRISS which employs a maximum likelihood method [1].

In this paper we propose a Bayesian method for prediction of survival, and discuss the results obtained on the US NTDB which includes about two million records of injured patients admitted to hospitals and emergency units [12]. The data include information about a patients age, gender, type and regions of injuries along with some clinical and background information about a patients state. The NTDB also includes information about TRISS prediction and outcome of care (alive or died) for each patient.

To test the proposed method, we selected a set of patients recorded in the NTDB with 1 to 20 injuries, and for which screening tests have been filled in; the number of such patients was about 0.5 million. The data have been divided into three injury groups. The first group includes the majority of patients which obtained 1 to 3 injuries; for this group the TRISS method has been designed. The second and third groups include a smaller number of patients with 4 to 10, and 11 to 20 injuries, respectively. The survival rate is lowest in the third group and highest in the first group.

We expect that the proposed method will outperform the TRISS method in the second and third groups, and will perform comparably in the first group. This research, however, is limited to records without missing values in screening tests, and we therefore do not generalize our method to the entire NTDB population. For evaluation of our method, a Bayesian Calculator of survival has been developed as a stand-alone application [32].

2. Bayesian Predictions

Typically, Bayesian inference assumes that there exist a number of models which are suitable to use for approximating the relationship between explana-

tory variables or predictors x and outcome variable y, which represent given data D. A model M is defined with parameters Θ that are fitted to data D, and so the goodness-of-fit of model M can be evaluated on the data D.

In the Bayesian context it is unnecessary to assume the existence of a true model. Instead of that, we assume that the average over suitable models $M_1, \ldots, M_k$ could result in an accurate approximation of the true relationship. The most efficient averaging over models is achieved within the Bayesian method [27]. Bayesian methods require to set a *prior* distribution of parameters Θ for a given model M, $f(\Theta|M)$, as well as a *likelihood* function $f(D|\Theta)$ to calculate a *predictive* posterior distribution $p(y|x, \Theta)$ according to Bayes's formula. When the desired distribution is calculated for a given model M, a number of problems have to be addressed. In this paper, we address a specific problem of Bayesian averaging over hierarchical models such as Decision Trees (DTs).

DT models are learnt from given data represented by a set of explanatory and dependent variables. The models learn to solve a problem using explanatory variables that make a distinguishable contribution to the problem. The variables make axis-parallel partitions of the data so that the user can interpret the DT models [5, 7, 13].

Figure 1 shows an example of a DT model consisting of two splitting nodes, s_1 and s_2, and three terminal nodes $t_1, \ldots, t_3$. The first node, s_1, called the root, splits the entire data into two disjoint subsets so that data samples from one subset fall into node s_2 via the left branch, and samples from the other subset fall into the terminal node t_2 via the right branch. The node s_2 further partitions the data samples which fall into the terminals t_2 or t_3 via the left and right branches. Finally one of the terminal nodes assigns the given input to one of the given classes.

The Bayesian method of averaging over DT models has been proposed in [11] and discussed in [13]. The method has been made computationally feasible with Markov Chain Monte Carlo (MCMC) simulation methods aimed at exploring a posterior density over DT model parameters by making random walk proposals. The desired density is approximated by drawing samples of the model parameters from areas of the parameter space (or areas of interest) that have high posterior density.

MCMC methods are intended to explore all possible areas of interest. However, posterior density is often multimodal and its detailed exploration cannot be achieved in a reasonable time. When this is the case, the MCMC approximation becomes inaccurate, and the Bayesian model averaging de-

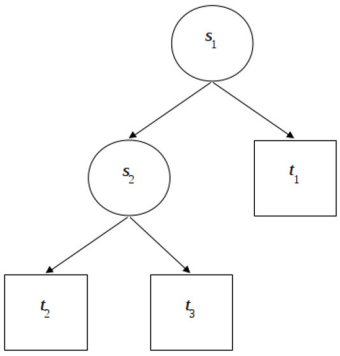

Figure 1: An example of DT model with splitting nodes s_1, s_2 and three terminal nodes $t_1, \ldots, t_3$

grades to the model selection, see e.g. [15].

Results of Bayesian averaging over DT models are dependent on prior distribution $p(\Theta)$ with which the user believes that Θ is the true DT model parameter. When the prior information is available, the averaging is mostly done over areas of interest with high posterior probability, and the averaging is likely to be accurate. However, when prior information is absent, the areas of possible interest cannot be specified and so may not be explored in detail. In such cases, the samples of Θ collected during MCMC approximation can disproportionally represent the posterior distribution $p(\Theta)$.

Particularly, we observed that when prior information on explanatory variables was absent, some DT models, namely samples of Θ, were over-represented in a DT ensemble collected during the MCMC simulation [19]. We evaluated the importance of these variables as frequencies of using them in the ensemble, and found that some of these variables have been used much less frequently. This allowed us to hypothesise that these variables made a weak contribution to the problem. We removed DT models which such variables from the ensemble, and observed a decrease in the uncertainty of the predictive density [20, 31].

In this paper we extend the Bayesian method of averaging over DT models to a large scale problem of predicting survival probabilities for patients recorded in the NTDB with 1 to 20 injuries. Additionally we explore the importance of the predictive variables for the prediction, and believe that this

information will be useful to optimize existing procedures of scoring injury severities.

3. MCMC Method of Averaging over DT Models

When averaging is made over DT models, the Bayesian formalism can be outlined as follows [11, 13]. We introduce a DT model with parameter Θ to be learnt from the data $\mathbf{D}$ represented by an m-dimensional input vector x and categorical outcome $y, y \in \{1, C\}$, where C is the number of categories inputs x can belong to. In our case the aim is to calculate the predictive distribution of survival probability, $p(y|x, \mathbf{D})$, for each patient.

Having defined models $M_1, \ldots, M_L$ with parameters $\Theta_1, \ldots, \Theta_L$, we can write the desired predictive distribution as an integral over the extended parameter vector $\Theta = (\Theta_1, \ldots, \Theta_L)$:

$$p(y|x, \mathbf{D}) = \int_{\Theta} p(y|x, \Theta) p(\Theta|\mathbf{D}) \mathrm{d}\Theta = \sum_{i=1}^{L} p(y|x, \Theta_i) p(\Theta_i|M_i, \mathbf{D}) p(M_i), \quad (1)$$

were $p(M_i)$ is the prior distribution of model M_i, $p(\Theta_i|M_i, \mathbf{D})$ is the posterior density of Θ_i given model M_i, and $p(y|x, \Theta_i)$ is the posterior predictive density given the parameters Θ_i.

This integral is analytically tractable only in cases when the distribution $p(\Theta|\mathbf{D})$ is given in an integrable form. However, in practice, we can only estimate this distribution by drawing N random samples $\Theta^{(1)}, \ldots, \Theta^{(N)}$ from the posterior distribution $p(\Theta|\mathbf{D})$, and then we can write:

$$p(y|x, \mathbf{D}) \approx \sum_{i=1}^{N} p(y|x, \Theta^{(i)}, \mathbf{D}) p(\Theta^{(i)}|\mathbf{D}) = \frac{1}{N} \sum_{i=1}^{N} p(y|x, \Theta^{(i)}, \mathbf{D}). \quad (2)$$

The above approximation is achieved with the MCMC method of simulation or stochastic integration. The approximation is achieved when a Markov chain becomes a random sequence with a stationary probability distribution. Then according to Eq. (2), we can draw the random samples $\Theta^{(i)}$ to calculate the desired predictive density.

In general, a DT model with k terminals consists of $(k-1)$ splitting nodes, $s_i, i = 1, \ldots, (k-1)$. The node s_i, has parameters including: the node position in the DT model, $s_i^p, p = 1, \ldots, (k-1)$, an input variable $s_i^v, v = 1, \ldots, m$, and a threshold s_i^q, where m is the number of variables

representing an input vector $x = (x_1, \ldots, x_m)$. The node s_i tests the vth variable against the threshold q and assigns the input x to the left branch if $x_v < q$, or to the right one otherwise. A terminal node t_i assigns the input x to class c with a probability $P_i^c, i = 1, \ldots, k$.

Consequently, a DT model is described by a vector of parameters, Θ, consisting of two parts. The first part includes the following parameters of nodes s_i: positions s_i^p, variables s_i^v, and thresholds s_i^q, $i = 1, \ldots, (k-1)$. The second part includes the probabilities $P_i^c, c = 1, \ldots, C$ for each terminal node $i, i = 1, \ldots, k$.

DT models whose nodes split data into two disjoint subsets are called binary. The number of possible configurations of binary DTs with k terminal nodes, S_k, is defined by the Catalan number:

$$S_k = \frac{1}{k+1}\binom{2k}{k}. \tag{3}$$

The number S_k grows exponentially with k and becomes very large for DTs with relatively small k. For example, for $k = 25, S_k$ becomes a number to the power of 12.

In practice, to explain data we need to induce DT models of a reasonable size; the size of a DT model is defined by the number of its terminal nodes, k. Oversized DT models are difficult to interpret, and moreover they are prone to overfit data.

The size of DT models is dependent on the number of data points, p_{min}, allowed to be in terminal nodes – setting a smaller p_{min} increases the size, while setting a greater p_{min} decreases the size. In most cases, prior information on the size of DT models is unavailable, and a suitable p_{min} has to be found empirically.

In practice, the size of DT models is unknown or can be given within a range. In such cases, areas of interest (high posterior density of parameters Θ), which have to be explored in Eq. 2, are of variable size, and MCMC has to be extended to Reversible Jump (RJ) proposed in [16].

Prior information about input variables, such as importance of variables $x_1, \ldots, x_m$, is also often unknown. In such cases, we can assign a variable v for the the node s_i to be drawn randomly from the uniform discrete distribution, $v \sim U(1, \ldots, m)$. Similarly, a threshold q can be drawn from the uniform discrete distribution, $q \sim U(min(x_v), max(x_v))$.

It has been shown that the above priors are sufficient in order to build and explore DT models of different configurations within the RJ MCMC method

[11, 13]. For binary DT models, the number of possible configurations, S_k, is defined by Eq. 3. From this equation, we see that the larger the k, the larger is the number S_k. So we expect that MCMC algorithm will explore possible DT configurations of size k with probabilities proportional to S_k.

The RJ MCMC method has been implemented for Bayesian averaging over DT models of variable size [11, 13, 31]. To explore DT models it has been proposed to use the *birth*, *death*, *change-split*, and *change-rule* moves made with Metropolis-Hastings (MH) sampler.

The first two, birth and death, moves were proposed to reversibly change the number of nodes in a DT model (or the dimensionality of the model parameter vector Θ). The third and fourth moves, change-split and change-rule, were aimed at changing the parameters Θ within a current dimensionality. The change-split move replaces a variable v in a chosen DT node s_i, while the change-rule move modifies a threshold q in node s_i.

The change-split moves are aimed at making large changes in the model parameters in order to potentially increase the chance of sampling from areas of interest. Such moves are intended to disrupt a long sequence of the posterior samples drawn from a local area of interest.

In contrast, the change-rule moves are aimed at making small changes in the parameters to let MCMC explore a surrounding area in detail. These moves are made more frequently than the others.

The MH sampler starts with a DT consisting of one splitting node whose parameter Θ is assigned within the predefined priors. Making the above moves, the sampler attempts to grow the DT model to a reasonable size by fitting its parameters Θ to the data. The fitness or likelihood of DT models is gradually increased and then becomes oscillatory around some value. This phase, named the *burn-in*, has to be preset sufficiently long in order to achieve a stationary distribution of the Markov chain. When the Markov chain becomes stationary, the samples of the posterior distribution are collected to approximate the desired predictive distribution – this phase is called *post burn-in*.

The above moves are made with the given proposal probabilities. Their values are dependent on the complexity of a classification problem – more complex problems require larger DT models. To grow such models, the proposal probabilities for the death and birth moves are set to larger values. In general, there is no guidance for setting proper parameters of the MH sampler, and their values have to be found empirically [11, 13, 31].

The proposed change is accepted according to Bayes' rule [13]. When the

birth or death move changes a dimensionality of a DT model, the acceptance rule needs to count a proposal ratio, R. This ratio is dependent on the number of possible configurations of DT models, S_k, and so we need to count R to keep the Markov chain reversible during the MCMC simulation. The reversibility is kept when the following condition is met:

$$q(\Theta|\Theta^p)p(\Theta^p) = q(\Theta^p|\Theta)p(\Theta), \tag{4}$$

where $q(\Theta^p|\Theta)$ is the conditional distributions of moving from the current parameter vector Θ to a proposed vector Θ^p, and $q(\Theta|\Theta^p)$ and $p(\Theta^p)$ are the densities of the reverse move.

4. Problems with the Metropolis-Hastings Algorithm

DT models are multilevel hierarchical structures, as shown in Figure 1. Nodes located at a lower hierarchical level are strongly dependent on the predecessor nodes located at upper level. In such hierarchical structures, changes proposed by the MH sampler can significantly redistribute data points falling into DT terminal nodes. The change made in a node close to the DT root is most influential on the distribution. The changes in terminal nodes can be so significant that the likelihood of the DT model is decreased – the closer the node is to the root, the more significant is the change in distribution of data points. In most cases such proposals are rejected. In contrast, a change proposed in a node close to DT terminals is most likely to be accepted as such a change will insignificantly redistribute data samples in the DT terminals. As a result, the MH sampler will only explore limited configurations of DT models [11, 13].

Another problem occurs when the MH algorithm aims to sample large DT models. When a DT model is small and consists of a small number of terminal nodes, the number of data samples falling into the nodes is expected to be much larger than the given minimal number of points, p_{min}. However, when a DT has grown large, the number of data points is decreased so that further partitions become unavailable. This means that birth moves cannot be made until a death move merges two terminal nodes into one node. As a result the MH algorithm will sample a series of DT models with similar distributions of data samples over terminal nodes. Such series affect the diversity of samples from the posterior distribution and, therefore, the accuracy of approximation of the predictive distribution [13, 29].

Another negative effect is that unavailable moves degrade the given proposal probabilities of birth and change moves. When a move is unavailable, the MH algorithm will repeat the current sample, which reduces the diversity of model mixing [13].

In most cases, the number p_{min} is found from experiments – complex problems typically require a small p_{min} to allow growth of large DT models. However, an inappropriately small p_{min} leads to excessive growth of DT models.

Growing a DT model, the MH algorithm makes birth moves and almost each birth move increases the likelihood of the model. The MH algorithm accepts these moves and the DT model grows rapidly. The growth of the model continues while the number of data samples in its terminal nodes exceeds p_{min} and the likelihood of a proposed model remains acceptable. During this period, the dimensionality of the DT model increases rapidly, and the sampler cannot explore the posterior within each dimensionality in detail. It is unlikely that samples will be drawn from areas of highest posterior density [11, 13].

The growth of DT models is typically monitored, and the modeller can reduce excessive growth by increasing p_{min} as well as by setting a smaller value of the proposal probability for the birth moves.

To mitigate the negative effect of fast growing DT models, Chipman et al [11] have proposed a restarting strategy. This strategy allows a DT model to grow within a limited period in multiple runs. The average over all models grown in these runs produces a better approximation accuracy when the duration of the growth period and the number of the runs are properly set.

A similar idea of restricting the growth of DT models has been proposed by Denison et al [13]. The growth is restricted within a given interval to allow the MH sampler to explore a model parameter space in detail. Both strategies require additional settings for the MH sampler, which have to be found experimentally.

As an alternative to the restricting strategies, the RJ MCMC method could be modified so as to reduce the number of replications of samples from the posterior density. In our previous work [29], we proposed a sweeping strategy aimed at reducing the number of unavailable moves.

For making a change-split move, the sweeping strategy assigns a new variable $x_v, v \sim U(1, m)$, and a threshold q:

$$q \sim U(a, b), \tag{5}$$

where $U(a, b)$ is a uniform distribution on the interval between $a = \min(x_{v,j})$ and $b = \max(x_{v,j})$ defined by N_p data points falling into the chosen node, where $j = 1, \ldots, N_p$.

For making change-rule moves, a new threshold q' is drawn from a restricted Gaussian distribution:

$$q' \sim N'(q, \sigma^2, a, b), \tag{6}$$

with mean q and given proposal variance σ^2 on the interval (a, b).

The proposed move can be made so that one or more terminal nodes in a DT model will contain fewer data points than p_{min}. If this happens in terminal nodes with a common parent node, these terminals are recombined into one terminal node, and the MH sampler counts such a move as a death move. If however there are two or more such terminals with different parents, the algorithm will assign the proposal unavailable in order to keep the reversibility of the Markov chain.

Similarly to a change move, a birth move assigns a new splitting node with parameters drawn from the given prior. A new splitting variable x_v is drawn from a uniform distribution, $v \sim U(1, m)$, and a new threshold q is assigned as described by Eq. 5.

In our experiments, we observed that a MH sampler using the above prior proposes fewer unavailable moves and, therefore, the sampler accepts fewer replications of a current parameter vector Θ. Taking this into account, we hypothesise that a reduced number of the replications collected during the MCMC simulation will improve the diversity of model mixing.

In support of this hypothesis, in our previous experiments [30] on the benchmark problems, we observed that the MH sampler using the above prior significantly reduced the dimensionality of parameter vector Θ as well as the uncertainty in estimates of predictive density. The above strategy, named *sweeping* in [29], is applied to the Markov chain in both burn-in and post burn-in phases.

As described in Section 3, a MH sampler makes the birth, death, and change moves. The sweeping strategy is implemented for the change move as follow.

1. Select a random splitting node $i \sim U(1, k-1)$ and read its variable v and threshold q
2. For change-split assign a new variable, $v' \sim U(1, m)$
3. For change-rule assign a new threshold q' defined by Eq. 6

4. Apply the proposed change to the DT and count the numbers of data points, p_i, falling into its terminal nodes
5. If $p_i \geq p_{min}$ for all $i = 1, \ldots, k$ terminal nodes, then go to Step 9
6. If else, find terminal nodes with $p_i < p_{min}$ and count their number n_0
7. If $n_0 == 1$, then apply the death move to the found node and go to Step 9
8. If $n_0 > 1$, then assign the proposal unavailable
9. Let MH sampler check acceptance of the proposal

In the following sections we explore the proposed strategy on a data set of patients from the NTDB. We attempt to improve the accuracy of estimates of predictive density.

5. Data

For this study, we selected a set of patients recorded in the NTDB with 1 to 20 injuries, who were alive on arrival at hospitals. TRISS predictions recorded in the NTDB were calculated with the MTOS coefficients. For comparison, we calculated predictions of survival probabilities with the regression coefficients updated on the NTDB; these coefficients were given in [34].

Table 1 shows the screening tests, variables $x_1, \ldots, x_{17}$, which were used for predicting survival. Variables x_1 (age), x_4 (blood pressure), and x_5 (respiration rate) are continuous, and the others are categorical; the output variable is the discharge status, $y = \{0, 1\}$. This table shows the ranges of the screening tests. We used these ranges to select a set of patients. After exclusion of missing and out-of-range values, the number of patients was 571,148.

The analysis of injuries of these patients showed that 67.3% of the whole population have 1 to 3 injuries, and 32.6% have obtained 4 to 20 injuries. The Table 2 shows the ratios and mortalities in the four groups of patients with 1 to 20, 1 to 3, 4 to 10, and 11 to 20 injuries. We can seen that the mortality is highest in the group with 11 to 20 injuries, and lowest in the group with 1 to 3 injuries.

The average values of the tests $x_1, \ldots, x_{17}$ in these groups are shown in Table 3. We can see that some values, such as Glasgow scores $x_6, \ldots, x_8$, change with an injury group.

It is obvious that the TRISS method will predict the survival most accurately for patients in the first injury group 1-3. On the contrary, for the

Table 1: Screening Tests with Ranges

Test	*Name*	*Range*
x_1	Age	0-99
x_2	Gender	0 female, 1 male
x_3	Injury type	0 penetrating, 1 blunt
x_4	Blood pressure	0-299
x_5	Respiration rate	0-59
x_6	GCS Eye	1-5
x_7	GCS Verbal	1-5
x_8	GCS Motor	1-6
x_9	Head severity	0-6
x_{10}	Face severity	0-6
x_{11}	Neck severity	0-6
x_{12}	Thorax severity	0-6
x_{13}	Spine severity	0-6
x_{14}	Abdomen severity	0-6
x_{15}	Upper extremity severity	0-6
x_{16}	Lower extremity severity	0-6
x_{17}	External severity	0-6
x_{18}	Discharge status	0 alive, 1 dead

other injury groups TRISS predictions will be less accurate because of the larger number of injuries a patient can obtain in the nine body regions.

Figure 2 shows the TRISS predictions and the observed survival probabilities for patients with 1 to 20 injuries. Here TRISS denotes the predictions with the NTDB regression coefficients, and TRISS^o denotes the predictions with the MTOS coefficients. We can see that the differences between the TRISS predictions and actual survival are progressively increased with the number of injuries.

Figure 3 shows the calibration curves of the TRISS model with the NTDB-based coefficients for patients in the four injury groups. We can see that the observed probabilities are significantly higher than the predicted values. The difference is largest for patients with a predicted survival between 0 to 0.7 in injury groups 1-20, 4-10, and 11-20.

The goodness-of-fit or calibration of prediction models can be evaluated

Table 2: Statistics of Injury Groups 1-20, 1-3, 4-10, and 11-20

	1-20	*1-3*	*4-10*	*11-20*
Population	100%	67.3%	29.9%	2.7%
Mortality	4.33%	2.69%	6.89%	16.87%

with the Hosmer-Lemeshow (HL) statistic as shown in the related literature [21, 3, 33]. The HL statistic is typically calculated for patients that fall into 10 intervals of survival probabilities. The smaller the value of HL statistic, the better is the calibration.

In our experiments the HL statistic values were calculated for each injury group as shown on the top of each plot in Figure 3. The calibration curve shown for injury group 1-3 shows unexpected fluctuations in the range of predicted survival between 0.3 and 0.6. These fluctuations can be caused by using the aggregated predictor ISS as discussed in [22].

6. Experiments

The set of patients described in the previous section was used for testing the Bayesian Decision Trees (BDT) we proposed for predicting survival. The proposed and TRISS methods are compared in terms of HL statistic, classification accuracy, and the area under the Receiver Operating Characteristic (ROC) curve, that are typically used in the related literature, see e.g. [21, 3].

The BDT was run with different settings, and the best results were obtained with the following settings. The proposal probabilities were set to 0.2, 0.2. 0.1, and 0.5 for the birth, death, change-split and change-rule moves, respectively. The proposal distribution was a Gaussian with zero mean. The numbers of samples for the burn-in and the post burn-in phases were 100,000 and 5,000, respectively. The minimal number of data samples allowed in DT terminals was set to 200.

Figure 4 shows the log likelihood, number of DT nodes, and distribution of DT sizes during the burn-in and the post burn-in phases. We see that the log likelihood and DT size became stable after 50,000 samples of burn-in. The collected DT models have grown on average to 150 nodes. The average acceptance rates were 0.4 for both burn-in and post burn-in phases, that lay in the optimal interval $(0.25, 0.5)$ according to [13].

Table 3: Mean Values of Screening Tests in Injury Groups

Test	*1-20*	*1-3*	*4-10*	*11-20*
x_1	39.29	39.79	38.28	37.70
x_2	0.66	0.64	0.70	0.69
x_3	0.88	0.86	0.92	0.97
x_4	134.57	136.10	132.35	120.83
x_5	19.09	19.29	18.79	17.29
x_6	3.72	3.84	3.52	2.86
x_7	4.55	4.73	4.27	3.33
x_8	5.60	5.77	5.32	4.32
x_9	0.93	0.61	1.49	2.49
x_{10}	0.35	0.19	0.65	1.16
x_{11}	0.03	0.02	0.05	0.11
x_{12}	0.66	0.33	1.25	2.47
x_{13}	0.38	0.22	0.65	1.36
x_{14}	0.38	0.24	0.62	1.26
x_{15}	0.56	0.40	0.86	1.40
x_{16}	0.85	0.72	1.04	1.79
x_{17}	0.11	0.09	0.14	0.18

The lower plots in Figure 3 show the calibration curves for the proposed method of predicting survival. We can see that the BDT predictions are much closer to the observed survival than the TRISS predictions shown in the upper plots in this figure. The values of the HL statistic shown in the plot titles were significantly smaller than those for the TRISS model.

The comparison of methods for predicting survival can be done in terms of the classification or discrimination accuracy which is calculated by assigning the outcome "alive" if a survival prediction is higher than 0.5, and "died" otherwise, see e.g. [21, 3]. Table 4 shows the classification accuracy (AC), sensitivity (SE), and specificity (SP) along with the area under ROC curve (AUC). We can observe that the AUC of the Bayesian method is slightly higher in all four groups, and for patients in groups 3-10 and 11-20 is higher by 1.2% than for the TRISS method. Figure 5 shows the ROC curves for both BDT and TRISS methods.

Figure 6 shows the uncertainty intervals in the four injury groups. We

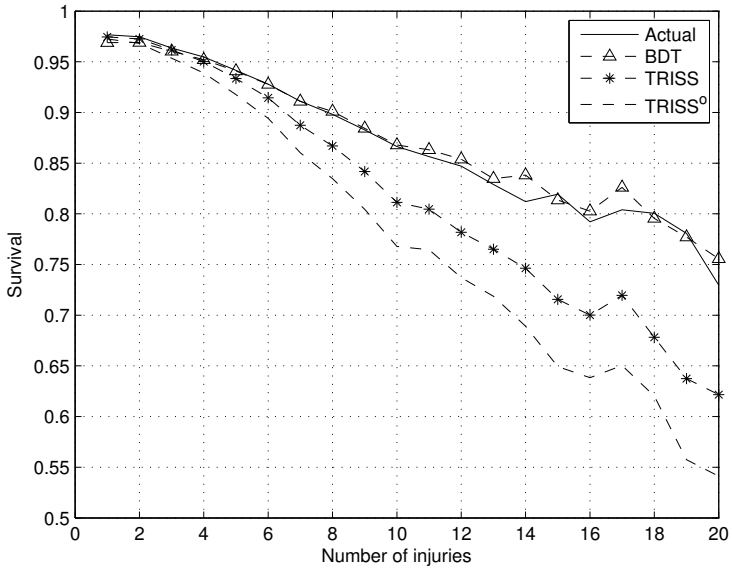

Figure 2: Observed and predicted survival probabilities for patients with different numbers of injuries.

can see that many of the TRISS predictions lie outside of the uncertainty intervals in all these groups. In contrast the BDT predictions are much close to the observed survival. The differences between the observed and predicted probabilities were estimated in terms of the weighted variance, v, shown in the plot titles.

Figure 2 shows the Bayesian (BDT) predictions versus the TRISS predictions and the actual survival probabilities for patients with injuries 1 to 20. We can see that the Bayesian predictions are much closer to the actual probabilities for patients with 4 and more injuries, while the TRISS predictions are mostly over-pessimistic.

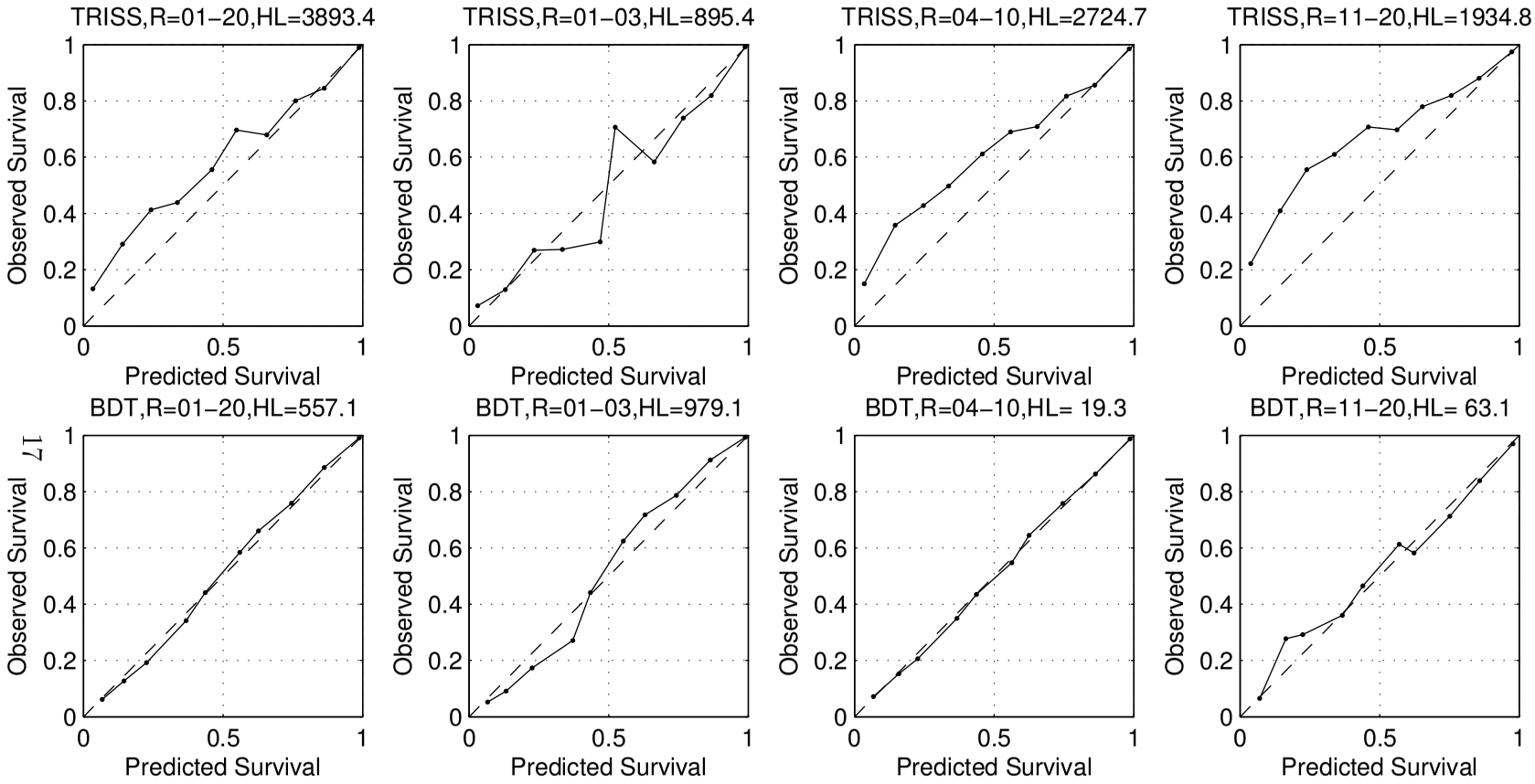

Figure 3: Calibration curves for TRISS and BDT model for patients in the four injury groups.

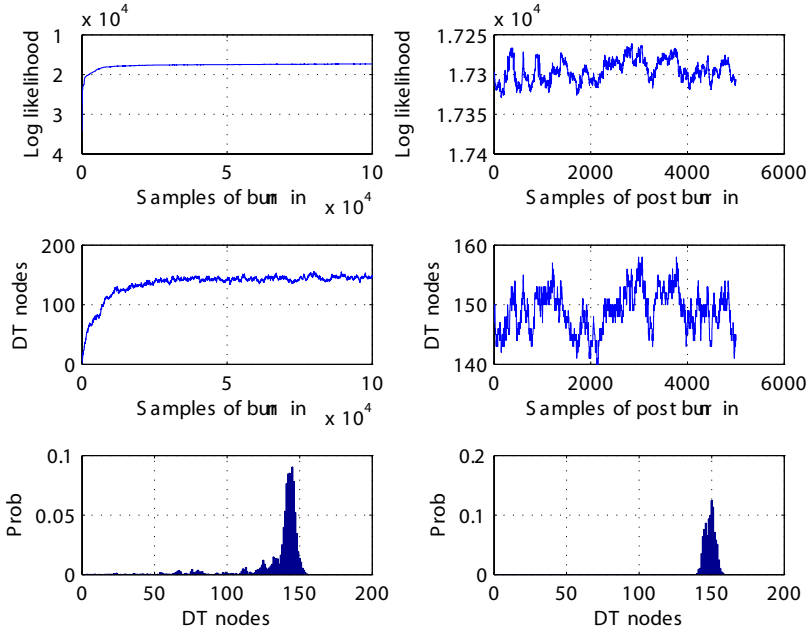

Figure 4: Likelihood (top), number of DT nodes (middle) and distribution of DT sizes (bottom) in the burn-in and post burn-in phases.

Table 4: Classification Accuracy for the Bayesian and TRISS Models in the four injury groups

	1-20		*1-3*		*4-10*		*11-20*	
	TRISS	BDT	TRISS	BDT	TRISS	BDT	TRISS	BDT
AC	0.968	0.971	0.983	0.984	0.944	0.952	0.838	0.875
SP	0.988	0.994	0.997	0.997	0.975	0.989	0.874	0.956
SE	0.528	0.474	0.489	0.504	0.532	0.447	0.664	0.475
AUC	0.948	**0.954**	0.950	**0.955**	0.932	**0.944**	0.882	**0.894**

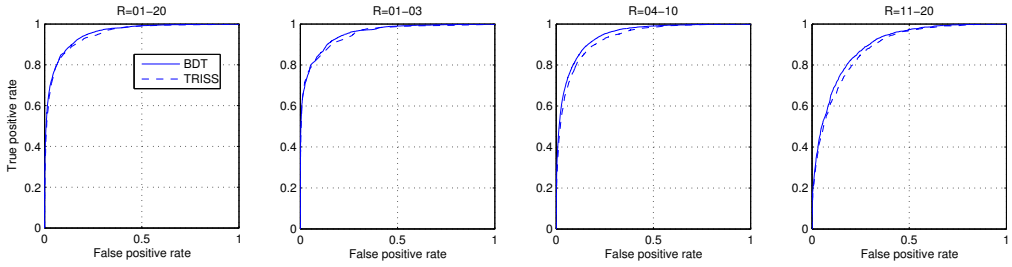

Figure 5: ROC curves for the TRISS and BDT models.

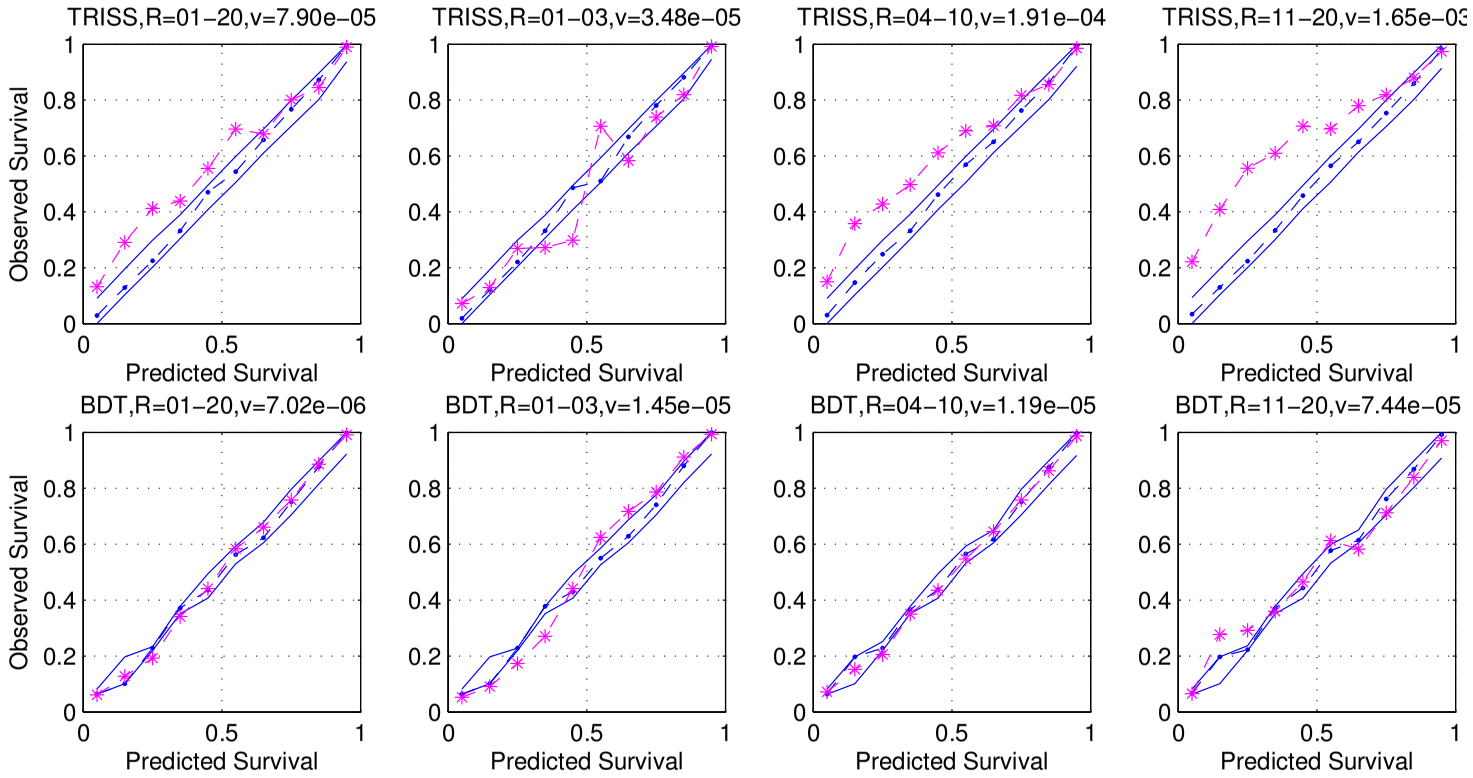

Figure 6: Uncertainty intervals for the BDT and TRISS models in the four groups.

7. Importance of Screening Tests

The ensemble of DT models collected during MCMC simulation allows us to estimate the contribution of the predictors $x_1, \ldots, x_{17}$ to the outcome. The importance of the predictors can be estimated in terms of frequencies (or posterior probabilities) of using them in the ensemble. These frequencies were calculated and shown in Table 5.

Table 5: Importance of Screening Tests

Rank	*Test*	*Name*	*Importance*
1	x_1	Age	0.156
2	x_{12}	Thorax severity	0.151
3	x_4	Blood pressure	0.109
4	x_{13}	Spine severity	0.107
5	x_9	Head severity	0.102
6	x_8	GCS Motor	0.068
7	x_7	GCS Verbal	0.060
8	x_{10}	Face severity	0.056
9	x_{16}	Lower extremity severity	0.055
10	x_{14}	Abdomen severity	0.034
11	x_3	Injury type	0.033
12	x_5	Respiration rate	0.032
13	x_{15}	Upper extremity severity	0.017
14	x_{17}	External severity	0.010
15	x_2	Gender	0.007
16	x_6	GCS Eye	0.002
17	x_{11}	Neck severity	0.001

We can see that the most important contribution is made by the predictor x_1 (Age), x_{12} (Thorax severity), and x_4 (Blood pressure). By contrast, the variables x_2 (Gender), x_6 (Glasgow Eye Coma Score), and x_{11} (Neck severity) are least important or redundant. Therefore their contribution can be insignificant for predicting the survival of patients with the proposed BDT method.

8. Bayesian Calculator of Survival Probabilities

For evaluating the proposed Bayesian method, we developed a Calculator for predicting survival and tested it on the data set described in Section 5. The Calculator allows the practitioner to predict survival probability for a given set of patient's screening tests.

It is important that the Calculator allows the practitioner to estimate the predictive probability density in order to assess the confidence intervals, which are associated with risk of making mistaken decisions. These estimates are made individually for each patient, whilst the TRISS method is unable to provide such estimates.

Figure 7 shows a screenshot of the calculator interface. The first column in the table Screening Tests shows the 17 screening tests that are described in Table 1. The second column shows the ranges of these tests. The third column displays values which the user can input or edit within the specified ranges.

The graph Predicted Probabilities of Survival displays the probabilities of survival for a patient with the given screening tests. Each of the predicted probabilities can be interpreted as a hypothesis which is tested on the data set in the context of Bayesian inference. The bars on the graph show the observed probabilities of these hypotheses. The estimates of the predictive density shown in the graph provide all the information required to calculate the confidence intervals.

Consider the example shown in Figure 7 recorded in the NTDB with a TRISS survival probability 0.680 and outcome "alive". For this patient, the Calculator predicts a probability of 0.582 within a 95% confidence interval 0.43 and 0.71. As this value exceeds 0.5, the predicted outcome is "alive". The locations and heights of the bars shown in the graph present the estimates of predictive density. All the bars on the left from the 0.5 mark on the x-axis represent low probabilities of survival associated with the outcome died whereas all the bars on the right represent high probabilities of survival. Observing these probabilities, the user can analyse the risk for this patient.

In this example, the sum over the first bars (on the left from 0.5) is smaller than the sum over the bars on the right. The substantial proportion of the former bars warns the user about a high death risk attached to this prediction. These bars make the predicted probability distribution wider and the uncertainty interval larger.

In addition to a predicted probability of survival, the user can analyse

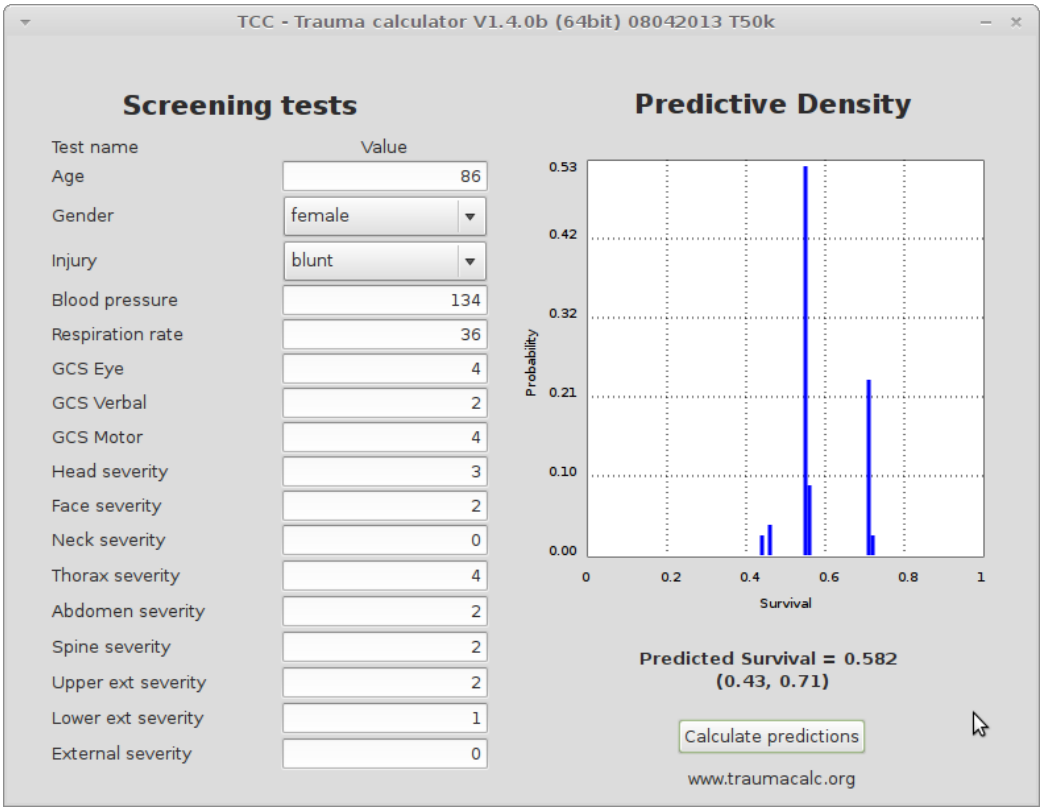

Figure 7: Bayesian Calculator Screenshot.

the confidence interval calculated for a given patient. The lengths of the intervals can be associated with the difficulty of treatment for patients – the larger the interval, the more difficult is the case.

The Calculator can be downloaded from a web page [32] to be run on a Windows XP 32-bit or Linux 64-bit machine. The Bayesian risk assessments are computationally expensive, and so a high performance machine with a 64-bit processor and 4 GB memory is recommended.

9. Discussion and Conclusion

We analysed the TRISS predictions for survival of patients recorded in the NTDB with 1 to 20 injuries and found that the goodness-of-fit and classification accuracy can be improved. The TRISS model cannot provide the

estimates of predictive probability density that are required to evaluate confidence intervals. The Injury Severity Score which is used as an aggregated predictor in TRISS model fluctuates unexplainably and may cause misleading predictions.

We explored the Bayesian Decision Tree models for predicting survival. The DT models are induced from data and capable of selecting important predictors.

Bayesian methods have been made computationally feasible by using MCMC simulation. However the results may be biased when samples are drawn from multiple areas of the posterior distribution of model parameters. We found that during the burn-in phase of the MCMC simulation, DT models tend to grow excessively, and that the existing MCMC strategies are unable to manage the growth efficiently in terms of diversity of model mixing.

We proposed a MCMC method capable of providing better conditions for detailed exploration of the posterior density during simulation. This method has been tested on a large set of patients recorded in the NTDB, and the results showed that the Bayesian Decision Tree model outperforms the TRISS model in terms of goodness-of-fit and classification accuracy.

We also showed that the ensemble of DT models collected during simulation allow practitioners to estimate the contribution of each screening test to the prediction. The importance of the tests was estimated as frequencies of their use by DT models.

The above results allows us to conclude that the proposed method can improve the accuracy of predicting survival for a patient with 4 to 20 injuries. The desired confidence intervals can be accurately estimated for each patient. Information about the importance of screening tests could be useful for cost analysis and for further improvement of the prediction accuracy.

[1] T. Bailey, R. Everson, J. Fieldsend, W. Krzanowski, D. Partridge, V. Schetinin, Representing classifier confidence in the safety critical domain an illustration from mortality prediction in trauma cases, Neural Computing and Applications 16 (2007) 1–10.

[2] D. Becalick, T. Coats, Comparison of artificial intelligence techniques with uktriss for estimating probability of survival after trauma. uk trauma and injury severity score, Journal of Trauma 51 (2001) 123–133.

[3] O. Bouamra, A. Wrotchford, S. Hollis, A. Vail, M. Woodford, F. Lecky,

A new approach to outcome prediction in trauma: A comparison with the triss model, Journal of Trauma 61 (2006) 701–710.

[4] C.R. Boyd, M.A. Tolson, W.S. Copes, Evaluating Trauma Care: The TRISS Method, Journal of Trauma 27 (1984) 370–378.

[5] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman and Hall, 1984.

[6] K. Brohi, TRISS - Overview and Desktop Calculator, http://www.trauma.org/index.php/main/article/387/, 2012.

[7] W. Buntine, Learning classification trees, Statistics and Computing 2 (1998) 6373.

[8] H.R. Champion, W.S. Copes, W.J. Sacco, M.M. Lawnick, S.L. Keast, C.F. Frey, The major trauma outcome study: Establishing national norms for trauma care, Journal of Trauma–injury Infection and Critical Care 30 (1990) 1356–1365.

[9] H.R. Champion, W.J. Sacco, W.S. Copes, Injury Severity Scoring Again, The Journal of Trauma: Injury, Infection, and Critical Care 38 (1995) 94–95.

[10] M. Chawda, F. Hildebrand, H. Pape, P. Giannoudis, Predicting outcome after multiple trauma: which scoring system?, Injury 35 (2004) 347–358.

[11] H. Chipman, E. George, R. McCullock, Bayesian CART model search, Journal of American Statistics 93 (1998) 935–960.

[12] Committee on Trauma. American College of Surgeons, NTDB Version 7.2, http://www.facs.org/trauma/ntdb/ntdbapp.html, 2007.

[13] D. Denison, C. Holmes, B. Mallick, A. Smith, Bayesian Methods for Nonlinear Classification and Regression, Wiley, 2002.

[14] S. DiRusso, T. Sullivan, C. Holly, S. Cuff, J. Savino, An artificial neural network as a model for prediction of survival in trauma patients: validation for a regional trauma area, Journal of Trauma 49 (2000) 220–223.

[15] P. Domingos, Bayesian averaging of classifiers and the overfitting problem, in: The 17th International Conference on Machine Learning, Morgan Kaufmann Publishers, 2000, pp. 223–230.

[16] P.J. Green, Reversible jump Markov chain Monte Carlo and Bayesian model determination, Biometrika 82 (1995) 711–732.

[17] J. Hilden, J.D. Habbema, B. Bjerregaard, The measurement of performance in probabilistic diagnosis. part ii: Trustworthiness of the exact values of the diagnostic probabilities, Methods of Information in Medicine 17 (1978) 227–237.

[18] A. Hunter, L. Kennedy, J. Henry, I. Ferguson, Application of neural networks and sensitivity analysis to improved prediction of trauma survival, Computer Methods and Programs in Biomedicine 62 (2000) 11–19.

[19] L. Jakaite, V. Schetinin, Feature selection for Bayesian evaluation of trauma death risk, in: The 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics, Springer, 2008, pp. 123–126.

[20] L. Jakaite, V. Schetinin, C. Maple, J. Schult, Bayesian decision trees for EEG assessment of newborn brain maturity, in: The 10th Annual Workshop on Computational Intelligence, UKCI 2010.

[21] P. Kilgo, J. Meredith, T. Osler, Incorporating recent advances to make the triss approach universally available, Journal of Trauma 60 (2006) 1002–1009.

[22] P. Kilgo, J. Meredith, T. Osler, Injury severity scoring and outcomes research, in: D.V. Feliciano, K.L. Mattox, E.E. Moore (Eds.), Trauma (6th ed), New York, McGraw-Hill, 2008, pp. 223–230.

[23] F. Millham, W. LaMorte, Factors associated with mortality in trauma: re-evaluation of the triss method using the national trauma data bank, Journal of Trauma 56 (2004) 1090–1096.

[24] J.S. Oakland, Statistical Process Control (5th edition), Butterworth-Heinemann, 2002.

[25] T. Osler, R. F.B., G. Badger, M. Healey, D. Vane, S. Shackford, A simple mathematical modification of triss markedly improves calibration, Journal of Trauma 53 (2002) 630–634.

[26] T. Osler, L. Glance, J. Buzas, D. Mukamel, J. Wagner, A. Dick, A trauma mortality prediction model based on the anatomic injury scale, Annals of Surgery 247 (2008) 1041–1048.

[27] C. Robert, G. Casella, Monte Carlo Statistical Methods, Springer Texts in Statistics, Springer, 2004.

[28] F. Rogers, T. Osler, M. Krasne, A. Rogers, E. Bradburn, J. Lee, D. Wu, N. McWilliams, M. Horst, Has triss become an anachronism? a comparison of mortality between the national trauma data bank and major trauma outcome study databases, Journal of Trauma and Acute Care Surgery 73 (2012) 326–331.

[29] V. Schetinin, J.E. Fieldsend, D. Partridge, W.J. Krzanowski, R.M. Everson, T.C. Bailey, The Bayesian decision tree technique with a sweeping strategy, in: The International Conference on Advances in Intelligent Systems, IEEE Computer Society, 2004.

[30] V. Schetinin, J.E. Fieldsend, D. Partridge, W.J. Krzanowski, R.M. Everson, T.C. Bailey, A. Hernandez, Comparison of the Bayesian and randomized decision tree ensembles within an uncertainty envelope technique, Journal of Mathematical Modelling and Algorithms 5 (2006) 397–416.

[31] V. Schetinin, L. Jakaite, Classification of newborn EEG maturity with Bayesian averaging over decision trees, Expert Systems with Applications 39 (2012) 9340–9347.

[32] V. Schetinin, L. Jakaite, J. Jakaitis, W. Krzanowski, Bayesian Calculator. Standalone application, http://traumacalc.org/bc2, 2012.

[33] E. Steyerberg, A. Vickers, N. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. Pencina, M. Kattan, Assessing the performance of prediction models: A framework for traditional and novel measures, Epidemiology 21 (2010) 128–138.

[34] Subcommittee on Trauma Registry Programs, American College of Surgeons Committee on Trauma, National trauma data bank: Reference manual, http://www.facs.org/trauma/ntdbmanual.pdf, 2004.